Genome analysis with gene expression microarrays

Mark Schena

Summary

Advances in biochemistry, chemistry and engineering have enabled the development of a new gene expression assay. This 'chip-based' approach utilizes microscopic arrays of cDNAs printed on glass as high-density hybridization targets. Fluorescent probe mixtures derived from total cellular messenger RNA (mRNA) hybridize to cognate elements on the array, allowing accurate measurement of the expression of the corresponding genes. Array densities of >1,000 cDNAs per cm^2 enable quantitative expression monitoring of a large number of genes in a single hybridization. A two-color fluorescence detection scheme allows rapid and simultaneous differential expression analysis of independent biological samples. Mass-produced microarrays provide a new tool for genome expression analysis that may revolutionize genetic dissection, drug discovery and human disease diagnostics.

Accepted
3 January 1996

Introduction

Genomics is an exciting new discipline in modern biology focusing on genome mapping, sequencing and analysis. Such studies promise to advance our understanding of biological systems and provide a wealth of fundamental and practical knowledge. The first complete DNA sequence of a free-living organism, that of the bacterium *H. influenzae*(1), paved the way for the acquisition of more elaborate genome sequences, including those of yeast and human. The expanding nucleotide and protein databases have necessitated more sophisticated methods of information handling; indeed, bioinformatics is a branch of computer science devoted entirely to managing and interpreting biological information(2).

The sequence of a genome can be viewed as a gene database for a given organism. Knowing the sequences of the genes, however, is only the first step in understanding the function of the genome. The intricate circuitry that governs growth, development, homeostasis, behavior and the onset of diseases is largely governed by the RNAs and proteins encoded by the cognate genes and the complex and dynamic interaction of the genes with the environment. The complexity of modeling simple gene regulatory networks, such as the lysis-lysogeny decision in the bacteriophage lambda(3), highlights the conceptual challenge presented by higher eukaryotes, which utilize hundreds of genetic pathways involving thousands of genes. A detailed conceptual view of gene regulatory circuitry in higher organisms will require extensive expression monitoring at the level of the whole genome. Biological analysis of this magnitude requires the development and implementation of sophisticated analytical methods.

The need for a practical, high-capacity gene expression assay prompted me to devise an alternative approach to gel- and filter-based methods employed in northern blots(4), S1 nuclease analysis(5), ribonuclease protection(6), primer extension(7) and differential plaque hybridization(8). The new methodology exploits recent advances in high-speed robotic printing of DNA samples(9,10), which allow the mass production of microscopic cDNA arrays for robust gene expression monitoring(10). The 'chip-based' approach, a term coined originally for oligonucleotide arrays(11-14), allows rapid quantitation of expression levels of many genes in parallel. Two-color fluorescence detection enables direct comparison of independent biological samples. Presented here is a discussion of the gene expression microarray methodology and how this technical advance might facilitate genome expression analysis and biological discovery, in both basic research and clinical settings.

Preparation of gene expression microarrays

Microarrays are printed on a glass surface using computer-controlled, high-speed robotics(9,10). The cDNAs to be arrayed are first amplified in a 96-well format using the polymerase chain reaction (PCR). Samples (nl) of the amplified and purified cDNAs are transferred from microtiter plates onto glass microscope slides using a robotic printhead (Fig.

ICSU Press 1996 pp. 427–431 BIO911

1). In preparation for monitoring gene expression, the cDNAs are linked chemically to the glass surface and denatured by heat treatment[9,10].

The source of cDNAs for the microarrays may include fully sequenced clones, collections of partially sequenced cDNAs known as expressed sequence tags (ESTs), or randomly chosen cDNAs from any library of interest. Amplifications can be carried out by using either gene-specific primers or primer pairs that recognize vectors sequences present in each clone. Common primers provide a significant advantage when amplifying clones in an automated system. A standard PCR reaction provides material sufficient to print more than 500 microarrays[9,10]. The printing speed allows 36 microarrays, each containing 1,000 array elements (cDNAs), to be fabricated automatically in about 5 hours. At a density of 1,000 cDNAs per cm^2, a 10 cm^2 microscope slide can provide specific hybridization targets for 10,000 genes.

Overview of the microarray assay

In principle, microarrays can be used to monitor expression in samples of any biological origin, including bacteria, fungi, higher plants and animals. In the protocol described in the first report[10], total mRNA was isolated from a biological sample and labeled using a single round of reverse transcription in the presence of fluorescent nucleotides (Fig. 2). The complex fluorescent probe mixture was then hybridized to a cDNA microarray, washed at high stringency and scanned with a laser. Fluorescence intensity at each position on the array provided an accurate measure of the expression of the cognate gene (Fig. 2).

Quantitative monitoring of gene expression

Gene expression microarray technology was developed using the small flowering plant *Arabidopsis thaliana* as a model system[10]. *Arabidopsis* offers many advantages for molecular genetic studies[15], including the fact that it has the smallest genome of any multicellular organism (Table 1), the complete DNA sequence is available for hundreds of *Arabidopsis* cDNAs, and large-scale sequencing efforts have succeeded in characterizing more than 20,000 expressed sequence tags (ESTs)[16,17]. Collections of cDNAs, in conjunction with the sequence database, expedite the preparation of microarrays and facilitate the interpretation of the expression data.

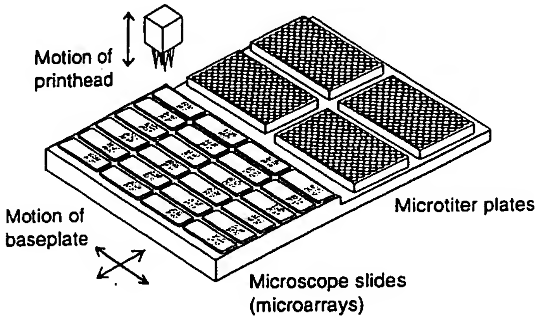

Fig. 1. Robotic arraying machine. The schematic diagram shows a microarraying device. A robotic printhead transfers a small sample of PCR-amplified cDNAs from microtiter plates to designated positions on a panel of microscope slides. Arrows indicate printhead and baseplate motions, both of which are computer-controlled.

A single 96-well microtiter plate of cDNAs was the source for the first gene expression arrays. The plate contained 96 samples, representing a total of 48 different clones, divided between adjacent wells to emphasize the reproducibility of the arraying and hybridization process. The 48 cDNAs included 45 cDNAs from *Arabidopsis* and one each from human, rat and yeast[10]. The 45 *Arabidopsis* cDNAs included 14 completely sequenced cDNAs and 31 ESTs[10]. The three non-plant cDNAs served as controls.

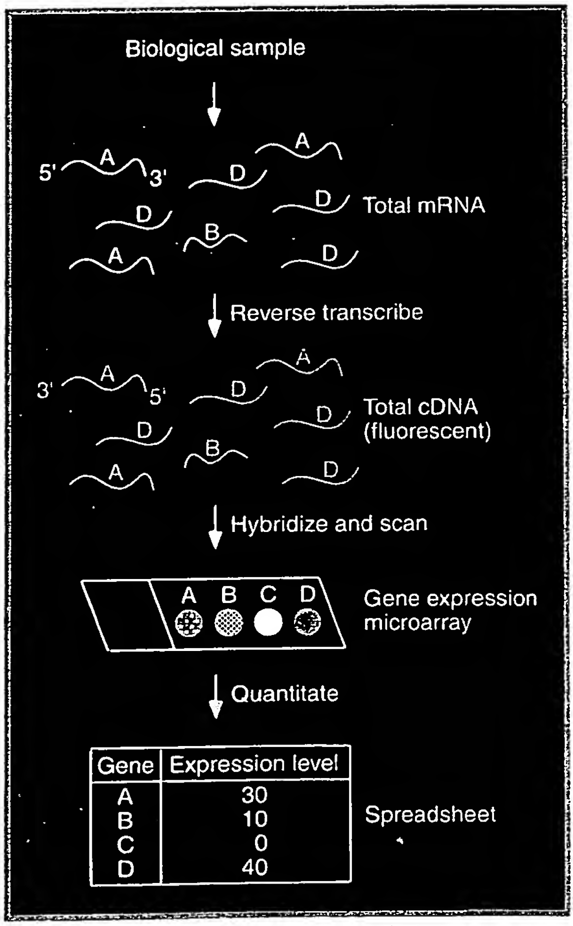

Fig. 2. Microarray assay for gene expression. The steps used to monitor gene expression on a cDNA microarray are outlined. Total mRNA is isolated from a biological sample, primed with oligo-dT and fluorescently labeled using a single round of reverse transcription in the presence of fluorescein-12-dCTP. The fluorescent probe is hybridized to a cDNA microarray containing specific hybridization targets, washed at high stringency and scanned for fluorescence emission following laser excitation. Measurement of fluorescence intensity allows quantitation of gene expression. The data are displayed on a spreadsheet.

Table 1. *Genome parameters for several eukaryotes*

Organism	Genome (nt)	Genes	ESTs
Saccharomyces cerevisiae (baker's yeast)	1.4×10^7	8,000	n.a.
Arabidopsis thaliana (thale cress)	0.7×10^8	25,000	22,593
Caenorhabditis elegans (nematode)	1.0×10^8	25,000	23,950
Homo sapiens (human)	3.0×10^9	100,000	307,214

Shown are approximate haploid genome sizes[23,24], gene numbers[23,24] and current EST availability[16,17] for several eukaryotic organisms. The EST database can be accessed on the world wide web (http://www.ncbi.nlm.nih.gov). Nucleotides, nt; not applicable, n.a.

2 µg of mRNA was labeled by incorporating fluorescein-12-dCTP during the reverse transcription step. The fluorescent cDNA mixture was hybridized to a microarray under a glass cover slip, washed and scanned for fluorescein emission following laser excitation. Hybridization with probe derived from root tissue, for example, revealed expression of many genes for which targets were arrayed, including the homeobox gene *KNAT1* (Fig. 3A; f3,4), the synaptobrevin gene *SAR1* (Fig. 3A; h5,6) and the cyclophilin gene *ROC1* (Fig. 3A; g7,8). Expression of other genes, such as the chlorophyll-binding protein gene (*CABl*), which is known to be highly repressed in root tissue, was not detected in the experiment (Fig. 3A; b1,2). Signals from known amounts of human acetylcholine receptor mRNA, added to the labeling reaction prior to reverse transcription, provided an internal calibration standard (Fig. 3A; a1,2).

Two-color fluorescence to monitor differential gene expression

Differential expression measurements are carried out using a simultaneous, two-color fluorescence hybridization scheme[9,10]. This strategy allows comparison of expression in virtually any two biological samples of interest. The samples are labeled independently with fluorescein- and lissamine-conjugated dCTP, respectively, mixed and hybridized to a single microarray. The array is then scanned at two wavelengths following independent excitation of the two fluors[9,10]. Direct comparison of the scanned images provides a rapid identification of genes whose expression is elevated or repressed in each sample[10]. Unlike conventional methods, the use of a single microarray for the measurements of differential expression avoids complications inherent in comparing results from independent hybridizations.

The first differential expression measurements[10] compared wild-type *Arabidopsis* plants with a transgenic line overexpressing a homeobox-leucine zipper gene known as *HAT4*[18]. 2 µg of mRNA from wild-type and *HAT4* transgenic plants were labeled with fluorescein and lissamine, respectively, mixed and hybridized to a single microarray. Comparison of the two scans revealed elevated *HAT4* expression in the transgenic line (Fig. 3C; e1,2) relative to wild-type plants (Fig. 3B; e1,2); the remaining 44 genes monitored on the array varied by less than fivefold between the two samples (Fig. 3B, C). Hybridization of fluorescein-labeled rat glucocorticoid receptor cDNA (Fig. 3B; c11,12) and lissamine-labeled yeast *TRP4* cDNA (Fig. 3C; h11,12) verified the presence of the negative control targets and the lack of optical cross talk between the two fluors.

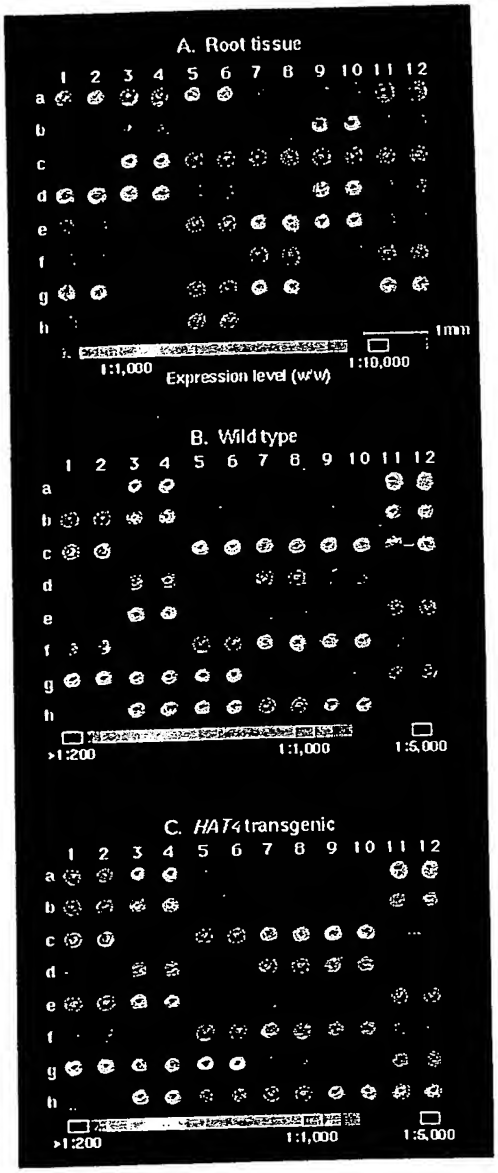

Fig. 3. Monitoring gene expression on a cDNA microarray. Images of gene expression microarrays are represented in a pseudocolor scale. Numbers and letters on the axes serve to identify the location of the gene-specific hybridization targets. Color bars denote gene expression levels. Scanned images are as follows: (A) mRNA from root tissue labeled with fluorescein and scanned for fluorescein emission; (B) mRNA from wild-type plants labeled with fluorescein and scanned for fluorescein emission; (C) mRNA from *HAT4* transgenic plants labeled with lissamine and scanned for lissamine emission. Note that two of the scans (B and C) were obtained from the same microarray scanned separately for fluorescein and lissamine, respectively.

Technical parameters of the microarray assay

The sensitivity of the microarray assay results from a small hybridization volume (2 μl), enabling a high probe concentration that exceeds those used in conventional northern blots by a factor of 10^5[4]. The current limit of detection for a single fluorescent species in a probe mixture derived from total mRNA is approx. 1:100,000 (w/w), which allows detection of nearly all cellular transcripts. Further improvements in the sensitivity of the assay might be accomplished by increasing the probe concentration and its specific activity, improving the detection system, reducing background fluorescence and optimizing the hybridization conditions.

The specificity of the assay is due to the use of relatively long hybridization targets (0.5-2.0 kb cDNAs) and stringent hybridization (65°C in 5×SSC + 0.1% SDS) and wash conditions (25°C in 0.1×SSC + 0.1% SDS). No detectable cross-hybridization is observed among even closely related members of gene families such as *HAT4* (Fig. 3; e1,2) and *HAT22* (Fig. 3; e9,10), which share greater than 70% nucleotide identity over an extended region[19]. It may be possible to further increase the specificity of the assay by modifying the coupling chemistry used to immobilize the target DNAs on the glass surface.

The use of a single round of reverse transcription for the labeling reaction produces a cDNA mixture that accurately reflects the mRNA population. Titration experiments indicate that the concentration of the probe is limiting for hybridization signals, even at the highest probe concentrations tested (2.0 μg/μl of a complex mixture). Target concentrations of 200 ng/cm^2 on the array surface allow linear hybridization kinetics over at least three orders of magnitude and quantitative detection of specific mRNAs representing as much as 2% of the total population. The robotic spotting method reproducibly delivers samples of a given cDNA that vary by less than twofold from element to element. In most cases, the level of gene expression measured on a microarray correlates with northern blots to within a factor of two[10]. Experiments are underway to explore the extent to which amplification during the mRNA labeling step alters quantitation. Probe amplification by reverse transcription-PCR (RT-PCR)[20] might allow for expression monitoring in minute samples of tissue such as human blood[10].

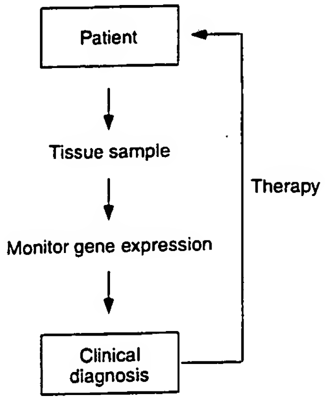

Fig. 4. Hypothetical flow chart for human disease diagnostics, showing one way in which cDNA microarrays might find use in molecular medicine. A tissue sample (e.g. blood or biopsy) from a patient with a physical or mental illness provides a source of mRNA for microarray-based gene expression monitoring[10]. Stereotyped changes in expression relative to a control sample might serve as a basis for clinical diagnosis and suggest an appropriate course of therapy[10]. Note that changes in expression could either trigger a disease state, or be the result of a given clinical condition.

Towards expression monitoring of whole genomes

The hierarchy in the control of gene expression in higher organisms provides an exceedingly complex problem for biological inquiry. For example, overexpression of the HAT4 transcription factor in plants triggers a cascade of developmental changes, including accelerated development and altered morphology and pigmentation[21]. The mechanistic details of action of HAT4 and many other transcription factors in higher plants and other organisms remain elusive. Though the consensus is that such regulators probably control the expression of large numbers of genes[22], identifying the upstream and downstream components of signaling pathways poses a difficult task. The capacity to monitor changes in gene expression at the whole genome level would expedite studies of this type by identifying groups of genes involved directly or indirectly in a given process.

The successful use of relatively simple gene expression microarrays sets the stage for more ambitious experiments. At a density of 1,000 cDNAs per cm^2, a single microarray of 8 cm^2 would provide gene-specific targets for all of the genes in yeast (Table 1) and five microarrays would allow whole-genome expression monitoring in *Arabidopsis* and *C. elegans* (Table 1). Though approx. 100,000 targets would be required for expression monitoring of the entire human genome, it is likely that a smaller number of targets, such as 10,000, would be adequate for most tissues. One microarray of human cDNAs may thus allow near-complete expression monitoring of peripheral blood, epithelium, liver, heart and other tissues and organs. The availability of large collections of ESTs from model organisms and human (Table 1) provide a rich source of targets for such arrays. Improvements in the DNA delivery system may allow array densities of 10^4-10^5 per cm^2.

Prospects for human biology

Microarrays could find a broad application in monitoring the differential expression in human genes, allowing changes in expression to be detected as a function of cell type, tissue source, physiological state or genetic background. Microarrays could also serve as a rapid method to identify changes in expression that accompany treatment of human cells with drugs, hormones, inhibitors, elicitors and a host of other small molecules.

It may be possible to utilize a microarray-based test for human disease diagnostics[10], whereby a sample of blood or a biopsy from a patient would provide a source of mRNA for monitoring gene expression (Fig. 4). Stereotyped changes in gene expression may accompany specific mental and physical illnesses[10] and thus provide diagnostic information and suggest particular courses of therapy (Fig. 4). Genes with altered expression patterns may be the disease genes *per se*, or genes whose expression is altered indirectly in afflicted individuals. Stereotyped expression patterns may also exist in subsets of the normal population including groups demarcated by race, gender, age and other differences. Extensive patient screening with gene expression microarrays might allow identification of informative patterns in each of these groups.

Acknowledgements

Thanks to Pat Brown, Dari Shalon, Renu Heller, Edgar Schreiber and Boris Zemelman for critical comments on the manuscript. This work was supported by Ronald W. Davis, who receives funding from the National Institutes of Health and the National Science Foundation.

References

1 **Fleischmann, R.D. *et al.*** (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

2 **Boguski, M.S.** (1994). Bioinformatics. *Curr. Opin. Genet. Dev.* **4**, 383-388.

3 **McAdams, H.H. and Shapiro, L.** (1995). Circuit simulation of genetic networks. *Science* **269**, 650-656.

4 **Alwine, J.C., Kemp, D.J. and Stark, G.R.** (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl Acad. Sci. USA* **74**, 5350-5354.

5 **Berk, A.J. and Sharp, P.A.** (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **12**, 721-732.

6 **Zinn, K., DiMaio, D. and Maniatis, T.** (1983). Identification of two distinct regulatory regions adjacent to the human β-interferon gene. *Cell* **34**, 865-879.

7 **McKnight, S.L. and Kingsbury, R.** (1982). Transcription control signals of a eukaryotic protein-encoding gene. *Science* **217**, 316-324.

8 **St John, T.P. and Davis, R. W.** (1979). Isolation of galactose-inducible DNA sequences from *Saccharomyces cerevisiae* by differential plaque filter hybridization. *Cell* **16**, 443-452.

9 **Shalon, D.** (1995). DNA micro arrays: a new tool for genetic analysis. PhD thesis, Stanford University.

10 **Schena, M., Shalon, D., Davis, R.W. and Brown, P.O.** (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470.

11 **Fodor, S.P.A., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. and Solas, D.** (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767-773.

12 **Southern, E.M., Maskos, U. and Elder, J.K.** (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: Evaluation using experimental models. *Genomics* **13**, 1008-1017.

13 **Guo, Z., Guilfoyle, R.A., Thiel, A.J., Wang, R. and Smith, L.M.** (1994). Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucl. Acids Res.* **22**, 5456-5465.

14 **Matson, R.S., Rampal, J., Pentoney, S.L., Jr., Anderson, P.D. and Coassin, P.** (1995). Biopolymer synthesis on polypropylene supports: Oligonucleotide arrays. *Anal. Biochem.* **224**, 110-116.

15 **Weigel, D. and Meyerowitz, E.M.** (1994). The ABCs of floral homeotic genes. *Cell* **78**, 203-209.

16 **Hofte, H. *et al.*** (1993). An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. *Plant J.* **4**, 1051-1061.

17 **Newman, T. *et al.*** (1994). Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. *Plant Physiol.* **106**, 1241-1255.

18 **Schena, M. and Davis, R.W.** (1992). HD-Zip proteins: Members of an *Arabidopsis* homeodomain protein superfamily. *Proc. Natl Acad. Sci. USA* **89**, 3894-3898.

19 **Schena, M. and Davis, R.W.** (1994). Structure of homeobox-leucine zipper genes suggests a model for the evolution of gene families. *Proc. Natl Acad. Sci. USA* **91**, 8393-8397.

20 **Kawasaki, E.S. *et al.*** (1988). Diagnosis of chronic myeloid and acute lymphocytic leukemias by detection of leukemia-specific mRNA sequences amplified *in vitro*. *Proc. Natl Acad. Sci. USA* **85**, 5698-5702.

21 **Schena, M., Lloyd, A.M. and Davis, R.W.** (1993). The *HAT4* gene of *Arabidopsis* encodes a developmental regulator. *Genes Dev.* **7**, 367-379.

22 **McKnight, S.L. and Yamamoto, K.R.** (ed.) (1992). *Transcriptional Regulation*, volumes 1 and 2. Cold Spring Harbor Laboratory Press, New York.

23 **Watson, J.D.** (1990). The Human Genome Project: Past, present and future. *Science* **248**, 44-49.

24 **Green, E.D. and Waterston, R.H.** (1991). The Human Genome Project, prospects and implications for clinical medicine. *J. Am. Med. Assoc.* **266**, 1966-1975.

Mark Schena is at the Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305-5307, USA.
E-mail: schena@cmgm.stanford.edu